SDS 383D Chapter 1: Preliminaries

Xizewen Han

March 11, 2019

Bayesian Inference in Simple Conjugate Families

(A)

Sol: For $i = 1, \dots, N$ such that $x_i \sim \text{Bernoulli}(w)$, we have

$$p(x_i|w) = w^{x_i}(1-w)^{1-x_i}$$

Therefore with prior $w \sim \text{Beta}(a, b)$, we have the posterior distribution

$$\begin{aligned} p(w|x_{1:N}) &\propto p(x_{1:N}|w) \cdot p(w) \\ &= \prod_{i=1}^{N} p(x_i|w) \cdot p(w) \\ &= w^{\sum x_i}(1-w)^{N-\sum x_i} \cdot \frac{1}{B(a,b)} w^{a-1}(1-w)^{b-1} \\ &\propto w^{\sum x_i + a - 1}(1-w)^{N-\sum x_i + b - 1} \end{aligned}$$

Note that the last expression is the kernel of $\text{Beta}(\sum_{i=1}^{N} x_i + a, N - \sum_{i=1}^{N} x_i + b)$ distribution, which by adding the normalizing term we'd have this as the posterior distribution. ■

(B)

Sol: For such transformation, which is one-to-one and onto, we have the inverse as

$$\begin{cases} x_1 = y_1 y_2, \\ x_2 = y_2 - y_1 y_2 \end{cases}$$

Therefore, we have the Jacobian as

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{bmatrix} = y_2 \geq 0$$

thus the joint distribution is

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= f_{X_1,X_2}\big(x_1(y_1,y_2),x_2(y_1,y_2)\big)|J| \\ &= f_{X_1}\big(x_1(y_1,y_2)\big)f_{X_2}\big(x_2(y_1,y_2)\big)|J| \\ &= \frac{1}{\Gamma(a_1)}(y_1y_2)^{a_1-1}e^{-y_1y_2}\cdot\frac{1}{\Gamma(a_2)}(y_2-y_1y_2)^{a_2-1}e^{-y_2+y_1y_2}\cdot y_2 \\ &= \underbrace{\frac{1}{\Gamma(a_1)\Gamma(a_2)}y_1^{a_1-1}(1-y_1)^{a_2-1}}_{g(y_1),\text{ kernel of Beta}(a_1,a_2)}\underbrace{y_2^{a_1+a_2-1}e^{-y_2}}_{h(y_2),\text{ kernel of Gamma}(a_1+a_2,1)} \end{aligned}$$

Since the joint distribution of Y_1 and Y_2 can be factored into two functions, $g(y_1)$ and $h(y_2)$, by Lemma 4.2.7 of Casella and Berger's Statistical Inference, as well as the form of each function noted above in the last expression, we conclude that Y_1 and Y_2 are independent random variables such that

$$Y_1 \sim \text{Beta}(a_1, a_2)$$

and

$$Y_2 \sim \text{Gamma}(a_1 + a_2, 1)$$

Based on such conclusion, we propose the following method to simulate a $\text{Beta}(a_1, a_2)$ random variable:

- 1st step: simulate from two Gamma random variables: $X_1 \sim \text{Gamma}(a_1, 1)$ and $X_2 \sim \text{Gamma}(a_2, 1)$
- 2nd step: calculate $Y_1 = \frac{X_1}{X_1+X_2}$

Such $Y_1 \sim \text{Beta}(a_1, a_2)$. ■

(C)

Sol: With $X_i \sim \mathcal{N}(\theta, \sigma^2)$ for $i = 1, \ldots, N$, in which θ is unknown and σ^2 is known, and $\theta \sim \mathcal{N}(m, v)$, we have the posterior distribution

$$\begin{aligned} p(\theta|x_{1:N}) &\propto p(\theta)\cdot p(x_{1:N}|\theta) \\ &= \frac{1}{\sqrt{2\pi v}}e^{-\frac{(\theta-m)^2}{2v}}\cdot\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\sum(x_i-\theta)^2}{2\sigma^2}} \\ &\propto \exp\Big(-\frac{(nv+\sigma^2)\theta^2-2\big(m\sigma^2+v(\sum x_i)\big)}{2\sigma^2 v}\Big) \end{aligned}$$

By noting the kernel, we recognize the posterior distribution as

$$\mathcal{N}\Big(\frac{m\sigma^2+v(\sum_{i=1}^{N}x_i)}{\sigma^2+nv},\frac{\sigma^2 v}{\sigma^2+nv}\Big)$$

The variance term can be re-written as

$$\frac{\sigma^2 v}{\sigma^2+nv} = \frac{1}{\frac{1}{v}+\frac{n}{\sigma^2}}$$

Note the denominator is the precision, and it's the sum of data precision and prior prevision – a good characteristic of posterior of normal-normal model. ■

(D)

Sol: With $X_i \sim \mathcal{N}(\theta, 1/w)$ for $i = 1, \ldots, N$, in which θ is known and precision w is unknown, and $w \sim \text{Gamma}(a, b)$, we have the posterior distribution

$$\begin{aligned} p(w|x_{1:N}) &\propto p(w) \cdot \prod_{i=1}^{N} p(x_i|\theta, w) \\ &\propto w^{a-1} e^{-\frac{1}{b}w} w^{\frac{N}{2}} e^{-\frac{\sum(x_i-\theta)^2}{2}w} \\ &= w^{\frac{N}{2}+a-1} e^{-\left(\frac{1}{b}+\frac{\sum(x_i-\theta)^2}{2}\right)w} \end{aligned}$$

By recognizing the kernel, we have the posterior distribution

$$w \sim \text{Gamma}\Big(\frac{N}{2} + a, \frac{1}{b} + \frac{\sum(x_i-\theta)^2}{2}\Big)$$

Since w is a re-parameterization of σ^2, we use change of variables directly on posterior pdf to derive the posterior of σ^2. We have the transformation $g(w) = \frac{1}{\sigma^2}$, which is one-to-one and onto, and

$$\frac{d}{d\sigma^2} g^{-1}(\sigma^2) = \frac{d}{d\sigma^2}(\sigma^2)^{-1} = -(\sigma^2)^{-2}$$

Therefore, for a general case $w \sim \text{Gamma}(a, b)$, we have

$$p(\sigma^2) = p\big(g^{-1}(\sigma^2)\big)\left|\frac{d}{d\sigma^2} g^{-1}(\sigma^2)\right| = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-(a-1)} e^{-\frac{b}{\sigma^2}} (\sigma^2)^{-2} = \frac{b^a}{\Gamma(a)}(\sigma^2)^{-a-1} e^{-\frac{b}{\sigma^2}}$$

i.e. $\sigma^2 \sim IG(a, b)$. Thus for our specific case, the posterior distribution of σ^2 is

$$\sigma^2 \sim IG\Big(\frac{N}{2} + a, \frac{1}{b} + \frac{\sum(x_i-\theta)^2}{2}\Big)$$

■

(E)

Sol: With known idiosyncratic variances for normal likelihood, the posterior density for unknown common mean θ is

$$\begin{aligned} p(\theta|x_{1:N}) &\propto \mathbb{P}(\theta) \cdot \prod_{i=1}^{N} p(x_i|\theta) \\ &= \frac{1}{\sqrt{2\pi v}} e^{-\frac{(\theta-m)^2}{2v}} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\sum \frac{(x_i-\theta)^2}{{\sigma_i}^2}} \\ &\propto e^{-\frac{1}{2}\left(\left(\frac{1}{v}+\sum \frac{1}{{\sigma_i}^2}\right)\theta^2 - 2\left(\frac{m}{v}+\sum \frac{x_i}{{\sigma_i}^2}\right)\theta\right)} \end{aligned}$$

We can thus directly write out the posterior distribution for θ as

$$\mathcal{N}\Big(\frac{\frac{1}{v}m + \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} x_i}{\frac{1}{v} + \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2}}, \frac{1}{v} + \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2}\Big)$$

notice that the mean term is expressed in the form as a weighted average of the observations x_i's and the prior mean m.

A later note: the weighted average part is more obvious if we work with precision instead of variance: denote prior precision as $w = \frac{1}{v}$ and idiosyncratic precision for each observation as $w_i = \frac{1}{\sigma_i{}^2}$, we have the posterior distribution in the form of

$$\mathcal{N}\Big(\frac{w \cdot m + \sum_{i=1}^N w_i \cdot x_i}{w + \sum_{i=1}^N w_i}, w + \sum_{i=1}^N w_i\Big)$$

■

(F)

Sol: Using the parameterization from the question sheet, we compute the marginal distribution of x as

$$\begin{aligned}
p(x) &= \int_w p(x|w)p(w)dw \\
&= \int_w (\frac{w}{2\pi})^{1/2} e^{-\frac{w}{2}(x-m)^2} \frac{(b/2)^{a/2}}{\Gamma(a/2)} w^{a/2-1} e^{-\frac{b}{2}w} \\
&= \frac{(b/2)^{a/2}}{\sqrt{2\pi}\Gamma(a/2)} \cdot \int_0^\infty \underbrace{w^{\frac{a+1}{2}-1} e^{-\left(\frac{b}{2}+\frac{(x-m)^2}{2}\right)w}}_{\text{kernel of Gamma}(\frac{a+1}{2}, \frac{b+(x-m)^2}{2})} dw \\
&= \frac{(b/2)^{a/2}}{\sqrt{2\pi}\Gamma(a/2)} \frac{\Gamma(\frac{a+1}{2})}{\left(\frac{b+(x-m)^2}{2}\right)^{\frac{a+1}{2}}} \cdot 1 \\
&= \frac{\Gamma(\frac{a+1}{2})}{\Gamma(\frac{a}{2})} \frac{1}{\sqrt{\pi b}} \Big(1 + \frac{(x-m)^2}{b}\Big)^{-\frac{a+1}{2}} \\
&= \frac{\Gamma(\frac{a+1}{2})}{\Gamma(\frac{a}{2})} \frac{1}{\sqrt{a\pi}(b/a)^{1/2}} \Big(\frac{a + \left(\frac{x-m}{(b/a)^{1/2}}\right)^2}{a}\Big)^{-\frac{a+1}{2}}
\end{aligned}$$

Note that the last expression is exactly the pdf of t location-scale distribution with $d = a$ degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$.

Later note: this question demonstrates a useful conclusion – if we want to sample from t-distribution, which has heavy tails, we don't have to sample directly; instead, we first sample precision from Gamma distribution, then sample observation from normal distribution given that precision. The resulting sample is a random variable with t-distribution. ■

The Multivariate Normal Distribution

(A)

Sol: For the two sub-problems, we do them by using the definition given by the question and directly applying matrix operation rules:

(1)

$$\begin{aligned}
cov(\boldsymbol{x}) &= \mathbb{E}\big((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\big) \\
&= \mathbb{E}\big((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}^T-\boldsymbol{\mu}^T)\big) \\
&= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T) \\
&= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) - \mathbb{E}(\boldsymbol{x})\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbb{E}(\boldsymbol{x}^T) + \boldsymbol{\mu}\boldsymbol{\mu}^T \\
&= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T \text{ (using the fact that } \mathbb{E}(\boldsymbol{x}) = \mu) \\
&= \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T
\end{aligned}$$

The last expression is what's asked to show. ■

(2)

Sol: The mean of $\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}$ is:

$$\mathbb{E}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) = \boldsymbol{A}\mathbb{E}(\boldsymbol{x}) + \mathbb{E}(\boldsymbol{b}) = \boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b}$$

Therefore, we have

$$\begin{aligned}
cov(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) &= \mathbb{E}\Big(\big((\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) - (\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b})\big)\big((\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) - (\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b})\big)^T\Big) \\
&= \mathbb{E}\Big(\big(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})\big)\big(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})\big)^T\Big) \\
&= \mathbb{E}\big(\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{A}^T\big) \\
&= \boldsymbol{A}\mathbb{E}\big((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\big)\boldsymbol{A}^T \\
&= \boldsymbol{A}cov(\boldsymbol{x})\boldsymbol{A}^T \text{ (the middle term is derived by the definition of covariance given by the question)}
\end{aligned}$$

The last expression is what's asked to show. ■

(B)

Sol: Since for all z_i's such that $i = 1, \ldots, p$, they are independent, the pdf of $\boldsymbol{z} = (z_1, \ldots, z_p)^T$, being the joint distribution of z_i's, is the multiplication of their individual densities:

$$\begin{aligned}
f_{\boldsymbol{Z}}(\boldsymbol{z}) &= \prod_{i=1}^{p} f(z_i) \\
&= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{{z_i}^2}{2}\Big) \\
&= \frac{1}{(2\pi)^{p/2}} \cdot \exp\big(-\frac{1}{2}\sum {z_i}^2\big) \\
&= \frac{1}{(2\pi)^{p/2}} \cdot \exp\big(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\big)
\end{aligned}$$

The mgf of $\boldsymbol{z}$ is given by

$$\begin{aligned}
\mathbb{E}\big(e^{\boldsymbol{t}^T\boldsymbol{z}}\big) &= \int_{z_1}\cdots\int_{z_p} e^{\boldsymbol{t}^T\boldsymbol{z}} f_{\boldsymbol{Z}}(\boldsymbol{z})dz_1\dots dz_p \\
&= \int_{z_1}\cdots\int_{z_p} e^{\sum t_i z_i} f_{\boldsymbol{Z}}(\boldsymbol{z})dz_1\dots dz_p \\
&= \int_{z_1}\cdots\int_{z_p} \prod_{i=1}^{p} e^{t_i z_i} \prod f_{Z_i}(z_i)dz_1\dots dz_p \\
&= \prod_{i=1}^{p}\int_{z_i} e^{t_i z_i} f_{Z_i}(z_i)dz_i \text{ (using the fact that } z_i\text{'s are independent)} \\
&= \prod_{i=1}^{p} \exp\big(mt_i + v{t_i}^2/2\big) \text{ (given by question)} \\
&= \prod_{i=1}^{p} \exp\big(0\cdot t_i + 1\cdot {t_i}^2/2\big) \text{ (given by question)} \\
&= \prod_{i=1}^{p} \exp\big({t_i}^2/2\big) \\
&= \exp\big(\frac{{t_i}^2}{2}\big) \\
&= \exp\big(\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}\big)
\end{aligned}$$

■

(C)

Sol: To show that if $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, its mgf is of the given form, we use the definition given in this problem: for a multivariate normal variable $\boldsymbol{x}$, $z=\boldsymbol{a}^T\boldsymbol{x}$ is univariate normal.

It's expectation is

$$\mathbb{E}(z)=\mathbb{E}(\boldsymbol{a}^T\boldsymbol{x})=\boldsymbol{a}^T\mathbb{E}(\boldsymbol{x})=\boldsymbol{a}^T\boldsymbol{\mu}$$

and

$$var(z)=cov(\boldsymbol{a}^T\boldsymbol{x})=\boldsymbol{a}^T cov(\boldsymbol{a}^T\boldsymbol{x})\boldsymbol{a}=\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}$$

We know the mgf of a univariate normal variable as

$$M_z(t)=(\boldsymbol{a}^T\boldsymbol{\mu})t+(\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})\frac{t^2}{2}=(t\boldsymbol{a})^T\boldsymbol{\mu}+\frac{(t\boldsymbol{a})^T\boldsymbol{\Sigma}(t\boldsymbol{a})}{2}$$

meanwhile

$$M_z(t)=\mathbb{E}(e^{zt})=\mathbb{E}(e^{(\boldsymbol{a}^T\boldsymbol{x})t})=\mathbb{E}(e^{(t\boldsymbol{a})^T\boldsymbol{x}})=M_{\boldsymbol{x}}(t\boldsymbol{a})$$

therefore, denoting $\boldsymbol{t}=t\boldsymbol{a}$, we have the mgf of multivariate normal variable $\boldsymbol{x}$ as

$$M_{\boldsymbol{x}}(\boldsymbol{t})=\boldsymbol{t}^T\boldsymbol{\mu}+\frac{\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}}{2}$$

For the other direction, i.e. with the given form of mgf, such variable is multivariate normal, we again use the definition given in this particular question s.t. every linear combination of a multivariate normal random variable is univariate normal:

for all vectors $\boldsymbol{a}$ such that $z = \boldsymbol{a}^T\boldsymbol{x}$, we have its moment generating function as

$$\begin{aligned} M_z(t) &= M_{\boldsymbol{a}^T\boldsymbol{x}}(t) \\ &= \mathbb{E}\big(e^{t\cdot(\boldsymbol{a}^T\boldsymbol{x})}\big) \\ &= \mathbb{E}\big(e^{(t\boldsymbol{a})^T\cdot\boldsymbol{x}}\big) \\ &= M_{\boldsymbol{x}}(t\boldsymbol{a}) \\ &= \exp\big((t\boldsymbol{a})^T\boldsymbol{\mu} + \frac{1}{2}(t\boldsymbol{a})^T\boldsymbol{\Sigma}(t\boldsymbol{a})\big) \\ &= \exp\big(t(\boldsymbol{a}^T\boldsymbol{\mu}) + \frac{1}{2}t^2(\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a})\big) \\ &= \exp\big(m\cdot t + v\cdot t^2/2\big) \end{aligned}$$

the last expression is exactly the form of the mgf of univariate normal distribution $\mathcal{N}(m, v)$, in which $m = \boldsymbol{a}^T\boldsymbol{\mu}$, $v = \boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a}$, i.e. z has a $\mathcal{N}(m, v)$ distribution. Therefore, according to the definition given by the question, $\boldsymbol{x}$ is multivariate normal. ■

(D)

Sol: The question defines

$$\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$$

in which

$$\boldsymbol{z} = (z_1, \ldots, z_p)^T$$

s.t. each z_i for $i = 1, \ldots, p$ are i.i.d. $\mathcal{N}(0, 1)$ distribution.

I want to solve this problem with two methods, with the intention of going through different parts of mathematical statistics knowledge as a review:

Method 1

Since z_i's are mutually independent, by Corollary 4.6.10 from Casella and Berger's *Statistical Inference* each of their linear combinations is univariate normal, thus $\boldsymbol{z}$ is a multivariate normal random variable. Its pdf can be derived by

$$p(\boldsymbol{z}) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} e^{-{z_i}^2/2} = \frac{1}{(2\pi)^{p/2}} e^{-\boldsymbol{z}^T\boldsymbol{z}/2}$$

which is the form of a standard multivariate normal variable's pdf – we just proved the claim/definition given from (B). We directly write out our result from (B): the mgf of $\boldsymbol{z}$ is

$$M_z(\boldsymbol{t}) = \exp\big(\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}\big)$$

Now we show $\boldsymbol{x}$ is a multivariate normal variable: using change of variable, we have

$$\boldsymbol{x} = g(\boldsymbol{z}) = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$$

since $\boldsymbol{L}$ is full rank, it's invertible, thus we have

$$\boldsymbol{z} = g^{-1}(\boldsymbol{x}) = \boldsymbol{L}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

Therefore we have the Jacobian

$$J = \Big|\frac{\partial \boldsymbol{r}(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big| = \Big|\frac{\partial \boldsymbol{g^{-1}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big| = \big|\boldsymbol{L}^{-1}\big| = \big|\boldsymbol{L}\big|^{-1}$$

The pdf of $\boldsymbol{x}$ is thus

$$\begin{aligned}
p(\boldsymbol{x}) &= \frac{1}{(2\pi)^{p/2}} e^{-(\boldsymbol{L^{-1}}(\boldsymbol{x}-\boldsymbol{\mu}))^{\boldsymbol{T}}(\boldsymbol{L^{-1}}(\boldsymbol{x}-\boldsymbol{\mu}))/2} \cdot |J| \\
&= (2\pi)^{-p/2}\big|\boldsymbol{L}\big|^{-1} \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{L}^{-1})^T(\boldsymbol{L}^{-1})(\boldsymbol{x}-\boldsymbol{\mu})\big) \\
&= (2\pi)^{-p/2}\big|\boldsymbol{L}\big|^{-1} \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{L}^{T})^{-1}(\boldsymbol{L}^{-1})(\boldsymbol{x}-\boldsymbol{\mu})\big) \\
&= (2\pi)^{-p/2}\big|\boldsymbol{L}\big|^{-1} \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{L}\boldsymbol{L}^{T})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)
\end{aligned}$$

Since $\boldsymbol{L}$ is full rank, $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{L}\boldsymbol{I}_{p\times p}\boldsymbol{L}^T$ is positive definite, in which $\boldsymbol{I}_{p\times p}$ is identity matrix (thus positive definite). Note that $p(\boldsymbol{x})$ is exactly the form of the pdf of $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^T)$ distribution.

Method 2

I also want to use (a simple manipulation on) mgf to solve this question:

Since we've already derived the mgf of $\boldsymbol{z}$ as a standard multivariate normal random variable, i.e.

$$M_z(\boldsymbol{t}) = \exp\big(\frac{1}{2}\boldsymbol{t}^T\boldsymbol{t}\big)$$

we can directly write out the mgf of $\boldsymbol{x}$ as

$$\begin{aligned}
M_{\boldsymbol{x}}(\boldsymbol{t}) &= \mathbb{E}(e^{t^T x}) \\
&= \mathbb{E}(e^{t^T(\boldsymbol{L}\boldsymbol{z}+\boldsymbol{\mu})}) \\
&= \exp\big(\boldsymbol{t^T\mu}\big)\mathbb{E}\big((\boldsymbol{L^T t})^T\boldsymbol{z}\big) \\
&= \exp\big(\boldsymbol{t^T\mu}\big)M_z(\boldsymbol{L^T t}) \\
&= \exp\big(\boldsymbol{t^T\mu}\big)\exp\big(\frac{1}{2}(\boldsymbol{L^T t})^T(\boldsymbol{L^T t})\big) \\
&= \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T L L^T t}\big)
\end{aligned}$$

thus by our conclusion from (C), a random variable with this form of mgf has a $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}})$ distribution.

A Later Note

We found out that the restriction such that $\boldsymbol{L}$ shall be a squared matrix can be relaxed – it can be a non-squared matrix, as long as it's full row rank (not that it's not full column rank!). We'd use this finding as one of our methods in solving Question (A) from **Conditionals and Marginals**.

To compute the expected value and covariance matrix of $\boldsymbol{x}$, we take the first and second derivative of mgf at $t = 0$, respectively:

$$\begin{aligned}
\mathbb{E}(\boldsymbol{x}) &= \frac{\partial M_{\boldsymbol{x}}(\boldsymbol{t})}{\partial \boldsymbol{t}}\Big|_{t=0} \\
&= \frac{\partial}{\partial \boldsymbol{t}} \int_{\boldsymbol{x}} \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{t}} \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{t}} \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big)(\boldsymbol{\mu} + \boldsymbol{LL^T t}) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \exp\big(\boldsymbol{0}\big)(\boldsymbol{\mu}) f(\boldsymbol{x}) d\boldsymbol{x} \\
&= \boldsymbol{\mu} \int_{\boldsymbol{x}} f(\boldsymbol{x}) d\boldsymbol{x} \\
&= \boldsymbol{\mu}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\boldsymbol{x^T x}) &= \frac{\partial^2 M_{\boldsymbol{x}}(\boldsymbol{t})}{\partial \boldsymbol{t}^T \boldsymbol{t}}\Big|_{t=0} \\
&= \frac{\partial}{\partial \boldsymbol{t}^T} \frac{\partial M_{\boldsymbol{x}}(\boldsymbol{t})}{\partial \boldsymbol{t}}\Big|_{t=0} \\
&= \frac{\partial}{\partial \boldsymbol{t}^T} \int_{\boldsymbol{x}} \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big)(\boldsymbol{\mu} + \boldsymbol{LL^T t}) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{t}^T} \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big)(\boldsymbol{\mu} + \boldsymbol{LL^T t}) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \Big(\exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big)(\boldsymbol{\mu} + \boldsymbol{LL^T t})^T (\boldsymbol{\mu} + \boldsymbol{LL^T t}) + \exp\big(\boldsymbol{t^T\mu} + \frac{1}{2}\boldsymbol{t^T LL^T t}\big)(\boldsymbol{LL^T})\Big) f(\boldsymbol{x}) d\boldsymbol{x}\Big|_{t=0} \\
&= \int_{\boldsymbol{x}} \Big(\exp\big(\boldsymbol{0}\big)(\boldsymbol{\mu})^T (\boldsymbol{\mu}) + \exp\big(\boldsymbol{0}\big)(\boldsymbol{LL^T})\Big) f(\boldsymbol{x}) d\boldsymbol{x} \\
&= (\boldsymbol{\mu^T\mu} + \boldsymbol{LL^T}) \int_{\boldsymbol{x}} f(\boldsymbol{x}) d\boldsymbol{x} \\
&= \boldsymbol{\mu^T\mu} + \boldsymbol{LL^T}
\end{aligned}$$

therefore

$$cov(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{x^T x}) - \mathbb{E}(\boldsymbol{x})^T \mathbb{E}(\boldsymbol{x}) = \boldsymbol{\mu^T\mu} + \boldsymbol{LL^T} - \boldsymbol{\mu}^T\boldsymbol{\mu} = \boldsymbol{LL^T}$$

■

(E)

Sol: For the "only if" direction, i.e. for multivariate normal variable $\boldsymbol{x}$, we could argue in the following manner: according to Casella and Berger's arguments under Theorem 2.3.12 in *Statistical Inference* about the uniqueness of Laplace transforms, each mgf can only have one pdf, i.e. one distinct distribution being

mapped to it. Therefore, the logic to solve this problem is that if we can find an affine transformation of independent univariate normal variables such that they have the same mgf as our desired multivariate normal variable's mgf, we can conclude that the transformation and the multivariate variable are from the same distribution, i.e. the multivariate normal variable can be expressed as an affine transformation of a standard multivariate normal variable.

We had from Question (C) such that for a $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ random variable $\boldsymbol{x}$, it has mgf

$$M_{\boldsymbol{x}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{\mu} + \frac{\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{\Sigma}\boldsymbol{t}}{2}\right)$$

Meanwhile, we've found in (D) s.t. an affine transformation $\boldsymbol{L}\boldsymbol{z}+\boldsymbol{\mu}$ to standard multivariate normal variable $\boldsymbol{z}$ has mgf

$$M_{\boldsymbol{L}\boldsymbol{z}+\boldsymbol{\mu}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{\mu} + \frac{\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}}\boldsymbol{t}}{2}\right)$$

Therefore, by simply observing the forms of these two mgf's we can find, by setting $\boldsymbol{L}$ s.t. $\boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}} = \boldsymbol{\Sigma}$, we'd have the same form of mgf. We can achieve this by performing Cholesky decomposition on $\boldsymbol{\Sigma}$: since $\boldsymbol{\Sigma}$ is positive definite, we can find $\boldsymbol{L}$ s.t. $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}}$.

We've thus found such affine transformation that $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$, in which $\boldsymbol{\mu}$ is the mean of $\boldsymbol{x}$ and $\boldsymbol{L}$ is the component of Cholesky decomposition of covariance matrix of $\boldsymbol{x}$.

With this conclusion, I propose the following algorithm to simulate a $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variable $\boldsymbol{x}$:

(1) simulate $\boldsymbol{z}$ from $\mathcal{MVN}(\boldsymbol{0}, \boldsymbol{I})$ distribution

(2) perform Cholesky decomposition on covariance matrix $\boldsymbol{\Sigma}$ of desired multivariate normal variable:

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}}$$

(3) compute $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$ – such $\boldsymbol{x}$ is a sample from $\mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. ■

(F)

Sol: We first note that a matrix quadratic form $Q(\boldsymbol{x} - \boldsymbol{\mu})$ has the form

$$Q(\boldsymbol{x} - \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T(A\boldsymbol{x} - \boldsymbol{\mu})$$

in which $\boldsymbol{A}$ is a symmetric matrix.

We've actually solved this problem in our solution for (D) under Method 1 – we applied change of variable on standard multivariate normal variable $\boldsymbol{z}$ to derive the pdf of $\boldsymbol{x} \sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^T)$. We briefly go over the steps again here:

The pdf of the standard multivariate normal variable $\boldsymbol{z}$ is

$$p(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}} e^{-\boldsymbol{z}^{\boldsymbol{T}}\boldsymbol{z}/2}$$

From (E) we have that $\boldsymbol{x} \sim \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be expressed by $\boldsymbol{z}$ as

$$\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$$

in which $\boldsymbol{LL^T} = \boldsymbol{\Sigma}$ is the Cholesky decompositon of $\boldsymbol{\Sigma}$.

Applying change of variable:

$$g(\boldsymbol{z}) = \boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$$

note that $\boldsymbol{L}$ here is invertible, thus we have

$$\boldsymbol{z} = g^{-1}(\boldsymbol{x}) = \boldsymbol{L}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

Therefore we have the Jacobian

$$J = \Big|\frac{\partial \boldsymbol{r}(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big| = \Big|\frac{\partial \boldsymbol{g^{-1}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\Big| = \big|\boldsymbol{L}^{-1}\big| = \big|\boldsymbol{L}\big|^{-1}$$

The pdf of $\boldsymbol{x}$ is thus

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{1}{(2\pi)^{p/2}} e^{-(\boldsymbol{L^{-1}}(\boldsymbol{x}-\boldsymbol{\mu}))^{\boldsymbol{T}}(\boldsymbol{L^{-1}}(\boldsymbol{x}-\boldsymbol{\mu}))/2} \cdot |J| \\ &= (2\pi)^{-p/2}\big|\boldsymbol{L}\big|^{-1} \exp\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{L}\boldsymbol{L}^T)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big) \end{aligned}$$

Note that $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$ which is a symmetric matrix, thus $(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{L}\boldsymbol{L}^T)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is a quadratic form. We thus can express the pdf as

$$p(\boldsymbol{x}) = C \cdot \exp\big(-\frac{1}{2}Q(\boldsymbol{x}-\boldsymbol{\mu})\big)$$

in which

$$C = (2\pi)^{-p/2}\big|\boldsymbol{L}\big|^{-1}$$

and

$$Q(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

■

(G)

Sol: To solve this problem, we first compute the mgf of $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x_1} + \boldsymbol{B}\boldsymbol{x_2}$, then use the fact that a mgf can uniquely determine a distribution to identify the distribution of $\boldsymbol{y}$ with corresponding parameters.

We have the mgf of $\boldsymbol{y}$ as

$$\begin{aligned}
M_{\boldsymbol{y}}(\boldsymbol{t}) &= \mathbb{E}(e^{\boldsymbol{t^T y}}) \\
&= \mathbb{E}(e^{\boldsymbol{t^T(Ax_1+Bx_2)}}) \\
&= \mathbb{E}(e^{\boldsymbol{t^T(Ax_1)}} e^{\boldsymbol{t^T(Bx_2)}}) \\
&= \mathbb{E}(e^{\boldsymbol{(A^Tt)^Tx_1}} e^{\boldsymbol{(B^Tt)^Tx_2}}) \\
&= \int_{\boldsymbol{x_1}} \int_{\boldsymbol{x_2}} e^{\boldsymbol{(A^Tt)^Tx_1}} e^{\boldsymbol{(B^Tt)^Tx_2}} f(\boldsymbol{x_1}, \boldsymbol{x_2}) d\boldsymbol{x_1} d\boldsymbol{x_2} \\
&= \int_{\boldsymbol{x_1}} \int_{\boldsymbol{x_2}} e^{\boldsymbol{(A^Tt)^Tx_1}} e^{\boldsymbol{(B^Tt)^Tx_2}} f(\boldsymbol{x_1}) f(\boldsymbol{x_2}) d\boldsymbol{x_1} d\boldsymbol{x_2} \text{ ($\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are independent)} \\
&= \int_{\boldsymbol{x_1}} e^{\boldsymbol{(A^Tt)^Tx_1}} f(\boldsymbol{x_1}) d\boldsymbol{x_1} \cdot \int_{\boldsymbol{x_2}} e^{\boldsymbol{(B^Tt)^Tx_2}} f(\boldsymbol{x_2}) d\boldsymbol{x_2} \\
&= M_{\boldsymbol{x_1}}(\boldsymbol{A^Tt}) \cdot M_{\boldsymbol{x_2}}(\boldsymbol{B^Tt}) \\
&= \exp\big(\boldsymbol{(A^Tt)^T\mu_1} + \frac{\boldsymbol{(A^Tt)^T\Sigma_1(A^Tt)}}{2}\big) \cdot \exp\big(\boldsymbol{B^Tt)^T\mu_2} + \frac{\boldsymbol{(B^Tt)^T\Sigma_2(B^Tt)}}{2}\big) \\
&= \exp\Big(\boldsymbol{t^T(A\mu_1 + B\mu_2)} + \frac{\boldsymbol{t^T(A\Sigma_1A^T + B\Sigma_2B^T)t}}{2}\Big)
\end{aligned}$$

This is almost the mgf of a multivariate normal variable, except that we need to justify that ($\boldsymbol{A\Sigma_1A^T + B\Sigma_2B^T}$) is positive definite so that it can be a covariance matrix:

Since both $\boldsymbol{A}$ and $\boldsymbol{B}$ have full column rank, and both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite since they are both covariance matrices, $\boldsymbol{A\Sigma_1A^T}$ and $\boldsymbol{B\Sigma_2B^T}$ are each positive definite. The sum of two positive definite matrices (with same dimension) is positive definite too: denote two p.d. matrices as $\boldsymbol{M}$ and $\boldsymbol{N}$, s.t.

$$\boldsymbol{u^TMu} > 0$$

and

$$\boldsymbol{u^TNu} > 0$$

for all nonzero vectors $\boldsymbol{u}$, then

$$\boldsymbol{u^T(M+N)u} = \boldsymbol{u^TMu} + \boldsymbol{u^TNu} > 0$$

Therefore, ($\boldsymbol{A\Sigma_1A^T + B\Sigma_2B^T}$) is positive definite, thus eligible to be a covariance matrix. We thus have

$$\boldsymbol{y} \sim \mathcal{MVN}\Big(\boldsymbol{A\mu_1 + B\mu_2}, \boldsymbol{A\Sigma_1A^T + B\Sigma_2B^T}\Big)$$

■

Conditionals and Marginals

(A)

Sol:

Method 1

This solution applies the affine transformation of $\boldsymbol{x}$: previously, we write $\boldsymbol{x}$ as an affine transformation of a standard multivariate normal random variable; this time, we directly apply affine transformation on $\boldsymbol{x}$.

Define $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ in which

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{k\times k} & \boldsymbol{0}_{(p-k)\times k} \end{bmatrix}$$

thus a k-by-p matrix with full row rank.

For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have the mgf of $\boldsymbol{y}$ as

$$M_{\boldsymbol{y}}(\boldsymbol{t}) = \mathbb{E}\big(e^{\boldsymbol{t}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{x}}\big) = \mathbb{E}\big(e^{(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{t})^{\boldsymbol{T}}\boldsymbol{x}}\big) = \exp\big((\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{t})^{\boldsymbol{T}}\boldsymbol{\mu}, \frac{(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{t})^{\boldsymbol{T}}\boldsymbol{\Sigma}(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{t})}{2}\big) = \exp\big(\boldsymbol{t}^{\boldsymbol{T}}(\boldsymbol{A}\boldsymbol{\mu}) + \frac{\boldsymbol{t}^{\boldsymbol{T}}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\boldsymbol{T}})\boldsymbol{t}}{2}\big)$$

Since $\boldsymbol{A}^{\boldsymbol{T}}$ has full column rank, given any non-zero vector $\boldsymbol{v}$ we have $\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{v}$ as a non-zero vector. Since $\boldsymbol{\Sigma}$ as a covariance matrix is positive definite, we have

$$\boldsymbol{v}^{\boldsymbol{T}}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\boldsymbol{T}})\boldsymbol{v} = (\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{v})^{\boldsymbol{T}}\boldsymbol{\Sigma}(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{v}) > 0$$

therefore $A\Sigma A^T$ is positive definite, thus eligible to be a covariance matrix. By our conclusion from Part (C) of previous question,

$$\boldsymbol{y} \sim \mathcal{MVN}(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\boldsymbol{T}})$$

Using the notation in our question, we have $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x_1}$, $\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu_1}$, $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\boldsymbol{T}} = \boldsymbol{\Sigma}_{11}$, therefore in other words,

$$\boldsymbol{x_1} \sim \mathcal{MVN}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}_{11})$$

Method 2

This method uses the affine transformation of standard multivariate normal variable that we studied from the previous question. I learnt this method from class and like it very much:

For $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$ in which

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{L}_{11} & \boldsymbol{L}_{12} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{bmatrix}$$

we have

$$\begin{aligned}
\boldsymbol{L}\cdot\boldsymbol{L}^{\boldsymbol{T}} &= \begin{bmatrix} \boldsymbol{L}_{11} & \boldsymbol{L}_{12} \\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{L}_{11}^T & \boldsymbol{L}_{21}^T \\ \boldsymbol{L}_{12}^T & \boldsymbol{L}_{22}^T \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{L}_{11}\cdot\boldsymbol{L}_{11}^T + \boldsymbol{L}_{12}\cdot\boldsymbol{L}_{12}^T & \boldsymbol{L}_{11}\cdot\boldsymbol{L}_{21}^T + \boldsymbol{L}_{12}\cdot\boldsymbol{L}_{22}^T \\ \boldsymbol{L}_{21}\cdot\boldsymbol{L}_{11}^T + \boldsymbol{L}_{22}\cdot\boldsymbol{L}_{12}^T & \boldsymbol{L}_{21}\cdot\boldsymbol{L}_{21}^T + \boldsymbol{L}_{22}\cdot\boldsymbol{L}_{22}^T \end{bmatrix} \\
&= \boldsymbol{\Sigma} \text{ (we have } \boldsymbol{L}\boldsymbol{L}^{\boldsymbol{T}} = \boldsymbol{\Sigma} \text{ from Part } (E) \text{ of previous question)} \\
&= \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}
\end{aligned}$$

Meanwhile,

$$\begin{aligned}\begin{bmatrix}\boldsymbol{x_1}\\ \boldsymbol{x_2}\end{bmatrix} &= \boldsymbol{x}\\ &= \boldsymbol{Lz}+\boldsymbol{\mu}\\ &= \begin{bmatrix}\boldsymbol{L}_{11} & \boldsymbol{L}_{12}\\ \boldsymbol{L}_{21} & \boldsymbol{L}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{z_1}\\ \boldsymbol{z_2}\end{bmatrix}+\begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{L}_{11}\boldsymbol{z_1}+\boldsymbol{L}_{12}\boldsymbol{z_2}\\ \boldsymbol{L}_{21}\boldsymbol{z_1}+\boldsymbol{L}_{22}\boldsymbol{z_2}\end{bmatrix}+\begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}\end{aligned}$$

in which

$$\boldsymbol{x_1} = \boldsymbol{L}_{11}\boldsymbol{z_1}+\boldsymbol{L}_{12}\boldsymbol{z_2}+\boldsymbol{\mu_1}$$

therefore, the mgf of $\boldsymbol{x_1}$ is

$$\begin{aligned}M_{\boldsymbol{x_1}}(\boldsymbol{t}) &= \mathbb{E}\big[\exp\big(\boldsymbol{t}^T(\boldsymbol{L}_{11}\boldsymbol{z_1}+\boldsymbol{L}_{12}\boldsymbol{z_2}+\boldsymbol{\mu_1})\big)\big]\\ &= \mathbb{E}[\exp\big(\boldsymbol{L}_{11}^T\boldsymbol{t}\big)^T\boldsymbol{z_1}]\cdot\mathbb{E}[\exp\big(\boldsymbol{L}_{12}^T\boldsymbol{t}\big)^T\boldsymbol{z_2}]\cdot\exp\big(\boldsymbol{t^T}\boldsymbol{\mu_1}\big)\\ &= \exp\big(\frac{1}{2}(\boldsymbol{L}_{11}^T\boldsymbol{t})^T(\boldsymbol{L}_{11}^T\boldsymbol{t})\big)\exp\big(\frac{1}{2}(\boldsymbol{L}_{12}^T\boldsymbol{t})^T(\boldsymbol{L}_{12}^T\boldsymbol{t})\big)\exp\big(\boldsymbol{t^T}\boldsymbol{\mu_1}\big)\\ &\quad(\boldsymbol{z_1}\text{ and }\boldsymbol{z_2}\text{ are independent standard multivariate normal variables})\\ &= \exp\big(\boldsymbol{t^T}\boldsymbol{\mu_1}+\frac{1}{2}\boldsymbol{t}^T(\boldsymbol{L}_{11}\cdot\boldsymbol{L}_{11}^T+\boldsymbol{L}_{12}\cdot\boldsymbol{L}_{12}^T)\boldsymbol{t}\big)\\ &= \exp\big(\boldsymbol{t^T}\boldsymbol{\mu_1}+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}_{11}\boldsymbol{t}\big)\end{aligned}$$

Here we are very tempted to write out the distribution of $\boldsymbol{x_1}$, but we do need to argue that $\boldsymbol{\Sigma}_{11}$ is positive definite. My current solution is to construct k-by-p matrix $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{I}_{k\times k} & \boldsymbol{0}_{(p-k)\times k}\end{bmatrix}$ as in **Method 1**, s.t. $\boldsymbol{\Sigma}_{11} = \boldsymbol{A\Sigma A^T}$ being positive definite is what we've showed before; but I haven't figured out a different way that circumvents such construction of $\boldsymbol{A}$.

After the above arguments, we can now write that

$$\boldsymbol{x_1} \sim \mathcal{MVN}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}_{11})$$

by the form of mgf. ■

(B)

Sol: Before we conduct the derivation, I just want to point out that $\boldsymbol{\Omega}$, just like $\boldsymbol{\Sigma}$, is symmetric:

Given any symmetric matrix $\boldsymbol{A}$, we have

$$\begin{aligned}&\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_{p\times p}\\ \Rightarrow&(\boldsymbol{A}^{-1}\boldsymbol{A})^T = \boldsymbol{A}^T(\boldsymbol{A}^{-1})^T = \boldsymbol{I}\\ \Rightarrow&\boldsymbol{A}(\boldsymbol{A}^{-1})^T = \boldsymbol{I}\\ \Rightarrow&(\boldsymbol{A}^{-1})^T = \boldsymbol{A}^{-1}\end{aligned}$$

Therefore, we might write $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \begin{bmatrix}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12}\\ \boldsymbol{\Omega}_{12}^T & \boldsymbol{\Omega}_{22}\end{bmatrix}$$

Since $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, we have

$$
\begin{aligned}
&\boldsymbol{\Omega}\boldsymbol{\Sigma} = \boldsymbol{I}_{p\times p} \\
\Rightarrow &\begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^T & \boldsymbol{\Omega}_{22} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_{k\times k} & \boldsymbol{0}_{k\times(p-k)} \\ \boldsymbol{0}_{(p-k)\times k} & \boldsymbol{I}_{(p-k)\times(p-k)} \end{bmatrix} \\
\Rightarrow &\begin{bmatrix} \boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{11} + \boldsymbol{\Omega}_{12}\boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{12} + \boldsymbol{\Omega}_{12}\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Omega}_{12}^T\boldsymbol{\Sigma}_{11} + \boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Omega}_{12}^T\boldsymbol{\Sigma}_{12} + \boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_{k\times k} & \boldsymbol{0}_{k\times(p-k)} \\ \boldsymbol{0}_{(p-k)\times k} & \boldsymbol{I}_{(p-k)\times(p-k)} \end{bmatrix} \\
\Rightarrow &\begin{cases} \boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{11} + \boldsymbol{\Omega}_{12}\boldsymbol{\Sigma}_{12}^T = \boldsymbol{I}_{k\times k} & (1) \\ \boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{12} + \boldsymbol{\Omega}_{12}\boldsymbol{\Sigma}_{22} = \boldsymbol{0}_{k\times(p-k)} & (2) \\ \boldsymbol{\Omega}_{12}^T\boldsymbol{\Sigma}_{11} + \boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{12}^T = \boldsymbol{0}_{(p-k)\times k} & (3) \\ \boldsymbol{\Omega}_{12}^T\boldsymbol{\Sigma}_{12} + \boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{22} = \boldsymbol{I}_{(p-k)\times(p-k)} & (4) \end{cases}
\end{aligned}
$$

From Equation (2) we have

$$
\begin{aligned}
\boldsymbol{\Omega}_{12}\boldsymbol{\Sigma}_{22} &= -\boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{12} \\
\Rightarrow \boldsymbol{\Omega}_{12} &= -\boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}
\end{aligned}
$$

Plugging it into Equation (1) we have

$$
\begin{aligned}
\boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{11} - \boldsymbol{\Omega}_{11}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T &= \boldsymbol{I}_{k\times k} \\
\Rightarrow \boldsymbol{\Omega}_{11}(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T) &= \boldsymbol{I}_{k\times k} \\
\Rightarrow \boldsymbol{\Omega}_{11} &= (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)^{-1}
\end{aligned}
$$

thus

$$
\boldsymbol{\Omega}_{12} = -(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}
$$

Meanwhile, with similar arguments, we have from Equation (3) s.t.

$$
\boldsymbol{\Omega}_{12}^T = -\boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}
$$

Plugging it into Equation (4) we have

$$
\begin{aligned}
-\boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} + \boldsymbol{\Omega}_{22}\boldsymbol{\Sigma}_{22} &= \boldsymbol{I}_{(p-k)\times(p-k)} \\
\Rightarrow \boldsymbol{\Omega}_{22}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) &= \boldsymbol{I}_{(p-k)\times(p-k)} \\
\Rightarrow \boldsymbol{\Omega}_{22} &= (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}
\end{aligned}
$$

and another expression for $\boldsymbol{\omega}_{12}^T$ (although we've derived $\boldsymbol{\omega}_{12}$ above):

$$
\boldsymbol{\Omega}_{12}^T = -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}
$$

We thus re-write $\boldsymbol{\Omega}$ as follows, in terms of blocks of $\boldsymbol{\Sigma}$:

$$
\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^T & \boldsymbol{\Omega}_{22} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)^{-1} & -(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{bmatrix}
$$

■

(C)

Sol: For this question, we first derive the conditional distribution in terms of the blocks of precision matrix (along with $\boldsymbol{x}$ and $\boldsymbol{\mu}$), then convert the parameters into terms of covariance matrix.

Now we begin the derivation:

to derive the conditional distribution of $\boldsymbol{x_1}$, we could utilize the relationship indicated by the following equation to derive the analytic form of the conditional pdf:

$$p(\boldsymbol{x}) = p(\boldsymbol{x_1}, \boldsymbol{x_2}) = p(\boldsymbol{x_1}|\boldsymbol{x_2}) \cdot p(\boldsymbol{x_2})$$

in which we could go through similar arguments as in (A) to derive the parameters of the marginal distribution of x_2, we don't need to do that – we can simply treat $\boldsymbol{x_2}$ as constant, and keep simplifying the pdf form until we only have terms including $\boldsymbol{x_1}$, while other terms would just be absorbed into the normalizing constant; and clearly, the marginal pdf of $\boldsymbol{x_2}$, $p(\boldsymbol{x_2})$, won't have the term $\boldsymbol{x_1}$.

Since the term $\boldsymbol{x_2}$ only appears in the exponent of the base of natural logarithm, we take the question's advice to work with the densities on a log scale, and gradually removing the terms that can be treated as constants. Again, notice that we are working with precision matrix during the derivation, instead of covariance matrix:

$$\begin{aligned}
\log p(\boldsymbol{x_1}|\boldsymbol{x_2}) &\propto \log p(\boldsymbol{x_1}, \boldsymbol{x_2}) \\
&\propto \log \exp\left(-\frac{1}{2}\Big(\begin{pmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \Big)^T \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^T & \boldsymbol{\Omega}_{22} \end{pmatrix} \Big(\begin{pmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \Big) \right) \\
&\propto \Big(\begin{pmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \Big)^T \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^T & \boldsymbol{\Omega}_{22} \end{pmatrix} \Big(\begin{pmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \Big) \\
&= (\boldsymbol{x_1} - \boldsymbol{\mu_1})^T \boldsymbol{\Omega}_{11} (\boldsymbol{x_1} - \boldsymbol{\mu_1}) + (\boldsymbol{x_2} - \boldsymbol{\mu_2})^T \boldsymbol{\Omega}_{12}^T (\boldsymbol{x_1} - \boldsymbol{\mu_1}) \\
&+ (\boldsymbol{x_1} - \boldsymbol{\mu_1})^T \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) + (\boldsymbol{x_2} - \boldsymbol{\mu_2})^T \boldsymbol{\Omega}_{22} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \\
&\propto (\boldsymbol{x_1} - \boldsymbol{\mu_1})^T \boldsymbol{\Omega}_{11} (\boldsymbol{x_1} - \boldsymbol{\mu_1}) + (\boldsymbol{x_2} - \boldsymbol{\mu_2})^T \boldsymbol{\Omega}_{12}^T (\boldsymbol{x_1} - \boldsymbol{\mu_1}) \\
&+ (\boldsymbol{x_1} - \boldsymbol{\mu_1})^T \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \\
&= (\boldsymbol{x_1} - \boldsymbol{\mu_1})^T \boldsymbol{\Omega}_{11} (\boldsymbol{x_1} - \boldsymbol{\mu_1}) + (\boldsymbol{x_2} - \boldsymbol{\mu_2})^T \boldsymbol{\Omega}_{12}^T \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{11} (\boldsymbol{x_1} - \boldsymbol{\mu_1}) \\
&+ (\boldsymbol{x_1} - \boldsymbol{\mu_1})^T \boldsymbol{\Omega}_{11} \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \\
&\propto \big((\boldsymbol{x_1} - \boldsymbol{\mu_1}) + \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \big)^T \boldsymbol{\Omega}_{11} \big((\boldsymbol{x_1} - \boldsymbol{\mu_1}) + \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \big)
\end{aligned}$$

Converting above result back from log scale, we have

$$\begin{aligned}
p(\boldsymbol{x_1}|\boldsymbol{x_2}) &= \exp\big(\log p(\boldsymbol{x_1}|\boldsymbol{x_2}) \big) \\
&\propto \exp\Big(-\frac{1}{2} \big((\boldsymbol{x_1} - \boldsymbol{\mu_1}) + \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \big)^T \boldsymbol{\Omega}_{11} \big((\boldsymbol{x_1} - \boldsymbol{\mu_1}) + \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}) \big) \Big)
\end{aligned}$$

which is exactly the form of the kernel of $\mathcal{MVN}\big(\boldsymbol{\mu_1} - \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12} (\boldsymbol{x_2} - \boldsymbol{\mu_2}), \boldsymbol{\Omega}_{11}^{-1}\big)$

We temporarily leave out the discussion on the positive-definiteness of $\boldsymbol{\Omega}_{11}^{-1}$, as it requires more linear-algebraic arguments; for now, I simply point out that it's symmetric:

$$(\boldsymbol{\Omega}_{11}^{-1})^T = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)^T = \boldsymbol{\Sigma}_{11}^T - \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}^{-1})^T \boldsymbol{\Sigma}_{12}^T = \boldsymbol{\Sigma}_{11}^T - \boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}^T)^{-1} \boldsymbol{\Sigma}_{12}^T = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T = \boldsymbol{\Omega}_{11}^{-1}$$

Therefore, in terms of (blocks of) covariance matrix, by our derivation in (B) we have we thus have the conditional distribution

$$\boldsymbol{x_1}|\boldsymbol{x_2} \sim \mathcal{MVN}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x_2} - \boldsymbol{\mu_2}), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)$$

■

Multiple Regression: Three Classical Principles for Inference

(A)

Sol: We show the derivation from each of the three principles individually as follows:

Least Squares

To find the minimum of the sum of squares w.r.t. $\boldsymbol{\beta}$, we first compute the critical point of the sum by solving for the $\boldsymbol{\beta}$ when its gradient equals to 0, then justifying the sum reaches its minimum by showing the Hessian matrix of $\boldsymbol{\beta}$ is positive definite:

For

$$\begin{aligned}\sum(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 &= \sum(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^T(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}) \\ &= \sum(y_i^T - \boldsymbol{\beta}^T\boldsymbol{x_i})(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}) \\ &= \sum(y_i^T y_i - 2y_i^T\boldsymbol{x_i}^T\boldsymbol{\beta} + \boldsymbol{\beta^T x_i x_i^T\beta})\end{aligned}$$

Note that although each y_i is a scalar, we treat them as a 1-by-1 matrix, thus has a meaning of being transposed. The gradient w.r.t. $\boldsymbol{\beta}$ is thus

$$\frac{\partial}{\partial\boldsymbol{\beta}}\sum(y_i^T y_i - 2y_i^T\boldsymbol{x_i}^T\boldsymbol{\beta} + \boldsymbol{\beta^T x_i x_i^T\beta}) = -2\big(\sum\boldsymbol{x_i}y_i\big) + 2\big(\sum\boldsymbol{x_i x_i^T\beta}\big)$$

Setting it equal to $\mathbf{0}$, we have

$$\begin{aligned}2\big(\sum\boldsymbol{x_i}y_i\big) + 2\big(\sum\boldsymbol{x_i x_i^T\beta^*}\big) &= \mathbf{0} \\ \Rightarrow \sum\boldsymbol{x_i x_i^T\beta^*} &= \sum\boldsymbol{x_i}y_i \qquad (1)\end{aligned}$$

It's tempting to leave it like that and derive for $\boldsymbol{\beta}^*$, but we rearrange the terms to make the summation sign disappear, since it would be much easier this way to argue that the Hessian matrix (which we'd compute later) is positive definite.

We define

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x_1}^T \\ \boldsymbol{x_2}^T \\ \vdots \\ \boldsymbol{x_n}^T\end{bmatrix} \Rightarrow \boldsymbol{X}^T = \begin{bmatrix}\boldsymbol{x_1} & \boldsymbol{x_2} & \dots & \boldsymbol{x_n}\end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix}y_1 \\ y_2 \\ \vdots \\ y_n\end{bmatrix}$$

we can see

$$\sum\boldsymbol{x_i x_i^T} = \boldsymbol{X^T X}, \sum\boldsymbol{x_i}y_i = \boldsymbol{X^T Y}$$

We shall directly claim that $\boldsymbol{X}$ has full column rank, because each of its column can be viewed as a feature, and generally all of the features shall be linearly independent – otherwise why bother having the extra feature that's linearly dependent on others, since we want the feature to help provide extra information in predicting $\boldsymbol{Y}$. Therefore, for n-by-p matrix $\boldsymbol{X}$, given any p-by-1 vector $\boldsymbol{u}$, we shall have $\boldsymbol{Xu}$ as a non-zero n-by-1 vector; or in other words, the null space of $\boldsymbol{X}$ is $\{\mathbf{0}\}$. We thus have

$$\boldsymbol{u^T X^T X u} = (\boldsymbol{Xu})^T(\boldsymbol{Xu}) > 0$$

i.e. $\boldsymbol{X^T X}$ is positive definite. Therefore, it is invertible, we can thus have $\boldsymbol{\beta}^* = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}$

To argue that such $\boldsymbol{\beta}^*$ gives us the minimum of $\sum(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2$, we compute the Hessian matrix of $\boldsymbol{\beta}$:

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \frac{\partial \sum (y_i - \boldsymbol{x_i^T \beta})^2}{\partial \boldsymbol{\beta}} = 2 \sum \boldsymbol{x_i x_i^T} = 2 \boldsymbol{X^T X}$$

which is clearly positive definite by our conclusion from above. Therefore, we'd achieve the minimum of $\sum (y_i - \boldsymbol{x_i^T \beta})^2$ at

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}$$

Maximum Likelihood under Gaussianity

For this question, we want to first find out the distribution for each y_i. After we do that, we plug the corresponding pdf into the function to be optimized. Eventually, we'd show that this optimization problem is actually equivalent to the one from **Least Squares**, thus can get the result for free by referring to the derivation from last section.

We first find out each y_i's distribution by computing its mgf: for each $y_i = \boldsymbol{x_i^T \beta} + \epsilon_i$, in which $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we have

$$\begin{aligned}
M_{y_i}(t) &= \mathbb{E}(e^{y_i t}) \\
&= \mathbb{E}\big(e^{(\boldsymbol{x_i^T \beta} + \epsilon_i)t}\big) \\
&= e^{(\boldsymbol{x_i^T \beta})t} \cdot \mathbb{E}(e^{\epsilon_i t}) \ (\boldsymbol{x_i^T \beta} \text{ is a constant term}) \\
&= e^{(\boldsymbol{x_i^T \beta})t} \cdot e^{\sigma^2 t^2/2} \\
&= \exp\big((\boldsymbol{x_i^T \beta})t + \sigma^2 t^2/2\big)
\end{aligned}$$

We recognize that the last expression is the mgf of a $\mathcal{N}\big(\boldsymbol{x_i^T \beta}, \sigma^2\big)$ random variable, thus is the distribution of each y_i.

Now using the fact that y_i's are independent, we have their joint distribution

$$p(y_{1,\ldots,n}) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \boldsymbol{x_i^T \beta})^2}{2\sigma^2}\right)$$

To maximize this joint likelihood, we have the following equivalent statements:

$$\begin{aligned}
& arg\max_{\boldsymbol{\beta}} \Big\{ \prod_{i=1}^{n} p(y_i) \Big\} \\
\equiv & arg\max_{\boldsymbol{\beta}} \Big\{ \log \prod_{i=1}^{n} p(y_i) \Big\} \\
\equiv & arg\max_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{n} \log p(y_i) \Big\} \\
\equiv & arg\max_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{n} \big(- \frac{(y_i - \boldsymbol{x_i^T \beta})^2}{2\sigma^2} \big) \Big\} \text{ (removing constants)} \\
\equiv & arg\min_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{n} \big((y_i - \boldsymbol{x_i^T \beta})^2 \big) \Big\} \text{ (flip negative sign to positive)}
\end{aligned}$$

The very last expression shows that, to find $\hat{\boldsymbol{\beta}}$ s.t. it maximizes the joint likelihood of y_i's is equivalent to find the $\hat{\boldsymbol{\beta}}$ that minimize the sum of squared errors. Therefore, the rest of derivation would just follow the same pattern as previous section, and of course we'd reach the same result.

Method of Moments

For this section, we first show that centering the data won't change the covariance: for our random variable $\boldsymbol{x}_i$, denote centered variable as $\boldsymbol{z}_i = \boldsymbol{x}_i - \mathbb{E}(\boldsymbol{x}_i)$, such that $\mathbb{E}(z_i) = \mathbb{E}(\boldsymbol{x}_i - \mathbb{E}(\boldsymbol{x}_i)) = \mathbf{0}$. For any variable $\boldsymbol{a}$, we have

$$cov(\boldsymbol{z_i}, \boldsymbol{a}) = \mathbb{E}\big[(\boldsymbol{z_i} - \mathbb{E}(\boldsymbol{z_i}))(\boldsymbol{a} - \mathbb{E}(\boldsymbol{a}))\big] = \mathbb{E}\big[(\boldsymbol{z_i})(\boldsymbol{a} - \mathbb{E}(\boldsymbol{a}))\big] = \mathbb{E}\big[(\boldsymbol{x_i} - \mathbb{E}(\boldsymbol{x_i}))(\boldsymbol{a} - \mathbb{E}(\boldsymbol{a}))\big] = cov(\boldsymbol{x_i}, \boldsymbol{a})$$

Therefore, we directly work with centered data. Assuming that our data is already centered, so that we can continue to use our previous notation $\boldsymbol{x_i}$. For each of the p predictors in random variable $\boldsymbol{x_i}$, we have the covariance between it and error term as

$$cov(x_{ij}, \epsilon_i) = \mathbb{E}\big[(x_{ij} - \mathbb{E}[x_{ij}])(\epsilon_i - \mathbb{E}[\epsilon_i])\big] = \mathbb{E}\big[(x_{ij})(\epsilon_i - \mathbb{E}[\epsilon_i])\big] = \mathbb{E}\big[x_{ij} \cdot \epsilon_i\big]$$

in which the last equation is true as we assume $\mathbb{E}[\epsilon_i] = 0$

Applying method of moments, we use sample moments to estimate the population moments. We have the sample covariance between j-th predictor and error term Q_j as

$$Q_j = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) \cdot \epsilon_i$$

in which the sample means are $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$. This equation is true because the data is centered.

By setting the sample covariance to 0, we have

$$\sum_{i=1}^{n} x_{ij} \cdot \epsilon_i = 0 \Rightarrow \sum_{i=1}^{n} x_{ij} \cdot (y_i - \boldsymbol{x_i^T}\boldsymbol{\beta_M}) = 0$$

in which $\boldsymbol{\beta_M}$ is the method of moments estimator of $\boldsymbol{\beta}$.

Since we have p predictors in total, and for each $j = 1, \ldots, p$ the same pattern follows, we can combine them into a system of p equations with number of unknowns being the dimension of $\boldsymbol{\beta_M}$, which is also p. We thus have the following equation:

$$\sum_{i=1}^{n} \boldsymbol{x_i} \cdot (y_i - \boldsymbol{x_i^T}\boldsymbol{\beta_M}) = 0 \Rightarrow \sum \boldsymbol{x_i} y_i - \sum \boldsymbol{x_i}\boldsymbol{x_i^T}\boldsymbol{\beta_M} = 0 \Rightarrow \sum \boldsymbol{x_i} y_i = \sum \boldsymbol{x_i}\boldsymbol{x_i^T}\boldsymbol{\beta_M}$$

Notice that the last equation is in exactly the same form as Equation (1) under **Least Squares** section. Therefore, the same conclusion follows, as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta_M}$. ■

(B)

Sol: We first derive the estimator with scalar w_i's as weights as follows:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{arg\,min}\Big\{ \sum w_i(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 \Big\} \\
&\equiv \underset{\boldsymbol{\beta}}{arg\,min} \begin{bmatrix} (y_1 - \boldsymbol{x_1}^T\boldsymbol{\beta})^T & (y_2 - \boldsymbol{x_2}^T\boldsymbol{\beta})^T & \dots & (y_n - \boldsymbol{x_n}^T\boldsymbol{\beta})^T \end{bmatrix} \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} \begin{bmatrix} y_1 - \boldsymbol{x_1}^T\boldsymbol{\beta} \\ y_2 - \boldsymbol{x_2}^T\boldsymbol{\beta} \\ \vdots \\ y_n - \boldsymbol{x_n}^T\boldsymbol{\beta} \end{bmatrix} \\
&\equiv \underset{\boldsymbol{\beta}}{arg\,min}\big(\boldsymbol{Y} - \boldsymbol{X\beta}\big)^T \boldsymbol{W}\big(\boldsymbol{Y} - \boldsymbol{X\beta}\big) \text{ (define the diagonal matrix from above as } \boldsymbol{W}_{n\times n}) \\
&\equiv \underset{\boldsymbol{\beta}}{arg\,min}\Big(\boldsymbol{Y^T W Y} - 2\boldsymbol{Y^T W X\beta} + \boldsymbol{\beta^T X^T W X\beta}\Big)
\end{aligned}$$

We apply similar arguments as those from **Least Squares** section below to derive $\hat{\boldsymbol{\beta}}$:

The gradient of $\boldsymbol{\beta}$ from the last expression above is

$$\frac{\partial}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X^T W Y} + 2\boldsymbol{X^T W X\beta}$$

and the Hessian matrix is

$$\mathcal{H}(\boldsymbol{\beta}) = 2\boldsymbol{X^T W X}$$

Given any non-zero vecto $\boldsymbol{u}$, we have

$$\boldsymbol{u^T X^T W X u} = (\boldsymbol{Xu})^T \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}(\boldsymbol{Xu}) = (\boldsymbol{W}^{1/2}\boldsymbol{Xu})^T(\boldsymbol{W}^{1/2}\boldsymbol{Xu})$$

As we pointed out before, since $\boldsymbol{X}$ has full column rank, $\boldsymbol{Xu}$ is a non-zero vector; and assuming none of the weights is 0, which is realistic (otherwise we don't even need this observation), we have non-zero vector $\boldsymbol{W}^{1/2}\boldsymbol{Xu}$. Therefore,

$$\boldsymbol{u^T X^T W X u} = (\boldsymbol{W}^{1/2}\boldsymbol{Xu})^T(\boldsymbol{W}^{1/2}\boldsymbol{Xu}) > 0$$

i.e. $\boldsymbol{X^T W X}$ is positive definite.

Therefore, by solving for $\boldsymbol{\beta}$ when gradient equals to 0 we have the critical point of the original summation to be minimized at $(\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W Y}$, and it gives the minimum of this summation due to the positive-definiteness of the Hessian matrix, i.e.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W Y}$$

We now derive the estimator with heteroscedastic Gaussian error:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{arg\,max}\big\{ \prod p(y_i) \big\} \\
&\equiv \underset{\boldsymbol{\beta}}{arg\,max}\big\{ \log \prod p(y_i) \big\} \\
&\equiv \underset{\boldsymbol{\beta}}{arg\,min}\Big\{ \sum \frac{1}{{\sigma_i}^2}(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 \Big\}
\end{aligned}$$

We notice from the last expression s.t. by setting each

$$\frac{1}{{\sigma_i}^2} = w_i$$

, the exact same arguments as our derivation above would follow. ■

Quantifying Uncertainty: Some Basic Frequentist Ideas

In linear Regression

(A)

Sol: We again use mgf to derive the sampling distribution of $\hat{\boldsymbol{\beta}}$ (not sure about the term "sampling"; here I'm deriving "the" distribution of $\hat{\boldsymbol{\beta}}$):

$$\begin{aligned}
M_{\hat{\boldsymbol{\beta}}}(\boldsymbol{t}) &= \mathbb{E}\big[e^{\boldsymbol{t^T}\hat{\boldsymbol{\beta}}}\big] \\
&= \mathbb{E}\Big[e^{\boldsymbol{t^T}\left((\boldsymbol{X^TX})^{-1}\boldsymbol{X^TY}\right)}\Big] \\
&= \mathbb{E}\Big[e^{\boldsymbol{t^T}\left((\boldsymbol{X^TX})^{-1}\boldsymbol{X^T}(\boldsymbol{X\beta}+\boldsymbol{\epsilon})\right)}\Big] \\
&= \mathbb{E}\Big[e^{\boldsymbol{t^T}\left((\boldsymbol{X^TX})^{-1}\boldsymbol{X^TX\beta}+(\boldsymbol{X^TX})^{-1}\boldsymbol{X^T\epsilon}\right)}\Big] \\
&= e^{\boldsymbol{t^T\beta}} \cdot \mathbb{E}\big[e^{\boldsymbol{t^T}\left((\boldsymbol{X^TX})^{-1}\boldsymbol{X^T\epsilon}\right)}\big] \\
&= e^{\boldsymbol{t^T\beta}} \cdot M_{\boldsymbol{\epsilon}}\big(\boldsymbol{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{t}\big) \\
&= \exp\big(\boldsymbol{t^T\beta}\big) \cdot \exp\Big(0 + \big(\boldsymbol{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{t}\big)^T\big(\boldsymbol{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{t}\big)\sigma^2/2\Big) \\
&= \exp\big(\boldsymbol{t^T\beta} + \boldsymbol{t^T}(\boldsymbol{X^TX})^{-1}\boldsymbol{t}\sigma^2/2\big)
\end{aligned}$$

Note that the last expression is the mgf of a $\mathcal{MVN}\big(\boldsymbol{\beta}, (X^TX)^{-1}\sigma^2\big)$ random variable. Since a mgf uniquely determines the distribution, we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{MVN}\big(\boldsymbol{\beta}, (X^TX)^{-1}\sigma^2\big)$$

■

(B)

Sol: To quantify the uncertainty of each of the p coefficients (including one for the intercept here), ideally we can simply read off the diagonal of $\hat{\boldsymbol{\beta}}$'s covariance matrix. However, this covariance matrix includes the term σ^2, which in our case we don't know its value. Therefore, we need to have an estimator of σ^2 from our sample data, and it would be nice if such estimator is unbiased.

Fortunately, we've been shown from last semester's Linear Models class such that, denoting residuals as $\hat{\boldsymbol{\epsilon}}$ such that

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$$

we'd have

$$\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}} \sim \sigma^2\mathcal{X}^2_{n-p}$$

Therefore, its expectation is

$$\mathbb{E}(\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}) = \sigma^2\mathbb{E}(\text{a } \mathcal{X}^2_{n-p} \text{ random variable}) = \sigma^2 \cdot (n-p)$$

We thus have an unbiased estimator of σ^2:

$$s^2 = \frac{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}{n-p}$$

We thus can estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$ by

$$S^2 = (\boldsymbol{X^T X})^{-1} \cdot s^2 = (\boldsymbol{X^T X})^{-1} \frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{n-p} = (\boldsymbol{X^T X})^{-1} \frac{(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T (\boldsymbol{Y} - \hat{\boldsymbol{Y}})}{n-p}$$

We've implemented such computation in R, and here's the code snippet:

```
# Fill in the blank
# compute residual
residual = y - x %*% betahat
res_T_res = t(residual) %*% residual
# calculate unbiased estimator of sigma squared
s_squared = (res_T_res/(dim(x)[1]-dim(x)[2]))[1,1]
# estimate covariance matrix of beta
betacov = solve(t(x) %*% x) * s_squared
# calculate standard error by hand
beta_se = sqrt(diag(betacov))
```

We found that our calculation of standard errors for each β_j exactly matches those from lm library:

```
> beta_se = sqrt(diag(betacov))
> beta_se_lm = sqrt(diag(betacovlm))
> beta_se
                      V5            V6            V7            V8            V9           V10
38.3286939914  0.0072511036  0.1741433372  0.0237693597  0.0692994207  0.1247136189  0.0003943697
          V11           V12           V13
 0.0147771925  0.1192916644  0.0048962536
> beta_se_lm
            x           xV5           xV6           xV7           xV8           xV9          xV10
38.3286939914  0.0072511036  0.1741433372  0.0237693597  0.0692994207  0.1247136189  0.0003943697
         xV11          xV12          xV13
 0.0147771925  0.1192916644  0.0048962536
>
```

My guess is that if this matches after this many decimal spaces, then lm's implementation might be exactly the same as the one we provided. ■

Propagating Uncertainty

(A)

Sol: Denoting the standard error of $f(\hat{\boldsymbol{\theta}})$ as $se\big(f(\hat{\boldsymbol{\theta}})\big)$, we have

$$\begin{aligned}
\big[se\big(f(\hat{\boldsymbol{\theta}})\big)\big]^2 &= var\big(f(\hat{\boldsymbol{\theta}})\big) \\
&= \mathbb{E}\big[\big(f(\hat{\boldsymbol{\theta}}) - \mathbb{E}f(\hat{\boldsymbol{\theta}})\big)^2\big] \\
&= \mathbb{E}\big[\big((\hat{\boldsymbol{\theta}}_{\mathbf{1}} + \hat{\boldsymbol{\theta}}_{\mathbf{2}}) - (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2)\big)^2\big] \\
&= \mathbb{E}\big[\big((\hat{\boldsymbol{\theta}}_{\mathbf{1}} - \boldsymbol{\theta}_1) + (\hat{\boldsymbol{\theta}}_{\mathbf{2}} - \boldsymbol{\theta}_2)\big)^2\big] \\
&= \mathbb{E}\big[(\hat{\boldsymbol{\theta}}_{\mathbf{1}} - \boldsymbol{\theta}_1)^2 + (\hat{\boldsymbol{\theta}}_{\mathbf{2}} - \boldsymbol{\theta}_2)^2 + 2(\hat{\boldsymbol{\theta}}_{\mathbf{1}} - \boldsymbol{\theta}_1)(\hat{\boldsymbol{\theta}}_{\mathbf{2}} - \boldsymbol{\theta}_2)\big] \\
&= var(\hat{\boldsymbol{\theta}}_{\mathbf{1}}) + var(\hat{\boldsymbol{\theta}}_{\mathbf{2}}) + 2 \cdot cov(\hat{\boldsymbol{\theta}}_{\mathbf{1}}, \hat{\boldsymbol{\theta}}_{\mathbf{2}}) \\
&= \hat{\boldsymbol{\Sigma}}_{11} + \hat{\boldsymbol{\Sigma}}_{22} + 2 \cdot \hat{\boldsymbol{\Sigma}}_{12}
\end{aligned}$$

thus the standard error is just taking the square root of our estimated variance of such linear function of estimated parameter.

We directly write out the square of standard error for $f(\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{p} \hat{\boldsymbol{\theta}}_j$:

$$\big[se\big(f(\hat{\boldsymbol{\theta}})\big)\big]^2 = var\big(f(\hat{\boldsymbol{\theta}})\big) = \sum_{j=1}^{p} \hat{\boldsymbol{\Sigma}}_{jj} + \sum_{1 \le i < j \le p} 2 \cdot \hat{\boldsymbol{\Sigma}}_{ij}$$

after which we take the square root to get the standard error.

Using matrix notation: define

$$\boldsymbol{J}_{p\times 1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^T$$

we have

$$var\big(f(\hat{\boldsymbol{\theta}})\big) = var\big(\sum_{j=1}^{p} \hat{\boldsymbol{\theta}}_j\big) = var\big(\boldsymbol{J}^T \hat{\boldsymbol{\theta}}\big) = \boldsymbol{J}^T var\big(\hat{\boldsymbol{\theta}}\big)\boldsymbol{J}$$

in which the (i,j)-th element in $var\big(\hat{\boldsymbol{\theta}}\big)$ is $\hat{\boldsymbol{\Sigma}}_{ij}$. ■

(B)

Sol: The first order Taylor approximation of $f(\hat{\boldsymbol{\theta}})$ around the unknown true parameter value $\boldsymbol{\theta}$ is

$$f(\hat{\boldsymbol{\theta}}) = f(\boldsymbol{\theta}) + \boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

Therefore, we approximate the variance of $f(\hat{\boldsymbol{\theta}})$ by

$$\begin{aligned} var\big(f(\hat{\boldsymbol{\theta}})\big) &= var\big(f(\boldsymbol{\theta}) + \boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\big) \\ &= var\big(\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\big) \\ &= \boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) var\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big)\boldsymbol{\nabla}^T_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \\ &= \boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) var\big(\hat{\boldsymbol{\theta}}\big)\boldsymbol{\nabla}^T_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \\ &= \boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta})\hat{\boldsymbol{\Sigma}}\boldsymbol{\nabla}^T_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \end{aligned}$$

in which the simplification of above steps uses the fact that true parameter $\boldsymbol{\theta}$ can be treated as a constant in the function of $\hat{\boldsymbol{\theta}}$.

Since the true parameter is unknown, we are not able to calculate the value $\boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta})$, therefore we use our estimated value $\hat{\boldsymbol{\theta}}$ in the place of $\boldsymbol{\theta}$ and have

$$var\big(f(\hat{\boldsymbol{\theta}})\big) = \boldsymbol{\nabla}_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\Sigma}}\boldsymbol{\nabla}^T_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}})$$

We observe three estimations:

- Taylor approximation of f
- approximation of gradient at unknown true $\boldsymbol{\theta}$ using the value of $\hat{\boldsymbol{\theta}}$
- approximation of covariance matrix by using square of standard error in place of σ^2

Each estimation could add extra deviation from true $var\big(f(\hat{\boldsymbol{\theta}})\big)$.

■